Recurrent Neural Network Properties and their Verification with Monte Carlo Techniques

Dmitry Vengertsev[1,2], Elena Sherman[1]
[1] Department of Computer Science, Boise State University
[2] Micron Technology, Technology and Product Development

“Characterizing the space of inputs that are processed correctly is central to the future of ML in adversarial settings, and **it will almost certainly be grounded in formal verification.”**

Ian Goodfellow

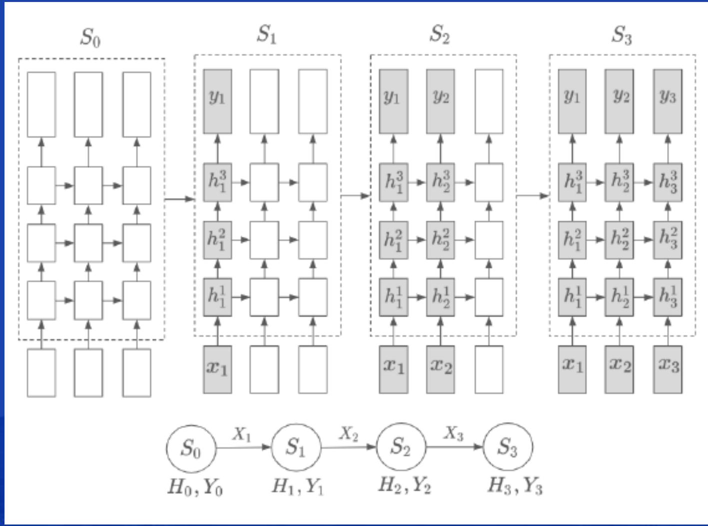
Key Research Contributions

- Define and formalize new state and temporal properties that are specific to RNNs
- Investigate whether Monte Carlo sampling is a suitable approach to verifying RNN models

1. We model RNN behavior as a Labeled Transition System (LTS) and uses LTL logic

RNN Behavioral model $M = (S, L, T, S_0)$ where:

- $S = \{(H, Y)_i\}_{i \in N}$ is a set of states that defined as tuple of hidden states and corresponding output value
- $L = \{X_i\}_{i \in N}$ is a finite set of labels that is based on the input vector X_i
- $T \subseteq S \times X \times S$ is a transition relation
- $S_0 = (H_0, Y_0)$ is an initial state



State Safety and Temporal Safety Properties

State Predicates

State predicates are functions over $S = (H, Y)$

- High confidence state predicate

$$Hi(a) : \bar{P}(Y) \geq a$$

- Low confidence state predicate

$$Lo(b) : \bar{P}(Y) \leq b$$

- Robustness state predicate

$$Ro(r, K) : \|Y_j - Y_i\| \leq K\|r\|$$

- Coverage state predicate

$$Cov(c, z) : \frac{\|H > z\|}{dim(H)} \geq c$$

State Safety Properties

- High-Confidence

$$GHi(a)$$

- Decisiveness

$$G(\neg Hi(a) \wedge Lo(b))$$

- Robustness

$$GRo(r, K)$$

- Coverage

$$GCov(c, z)$$

Temporal Safety Properties

- Long-term Relationship

$$G\eta(u, v, a, d)$$

$$\eta : \eta_n(u, a) \wedge \eta_{\rho(n)}(v, b)$$

$$\eta_n(u,a) : (Hi(a)(\neg Hi(a))^*)^u$$

$$\eta_{\rho(n)}(v, d) : (Hi(d)(\neg Hi(d))^*)^v$$

- Memorization

$$G\mu(q, e)$$

$$\mu : ((\neg Hi(e))^* Hi(e)(\neg Hi(e))^*)^q$$

Results

- Property satisfaction rates for both nextchar RNN models are not sufficient to be considered safe

- Comparing to the entire state space Monte Carlo sampling is efficient for estimating properties of RNN models

- The state safety properties are more efficiently checked than the temporal safety properties

Property	Notation	Ground Truth		Samples		ρ convergence	
		M1	*M2*	*M1*	*M2*	*M1*	M2
High Confidence	$GHi(a)$	**29.2**	20.6	5,371(0.9%)	3,055(0.5%)	8.2e-05	1.1e-04
Decisiveness	$G(\neg Hi(a) \wedge Lo(b))$	22.8	**26.0**	5,343(0.9%)	3,833(0.6%)	5.8e-05	1.1e-04
Robustness	$GRo(r, K)$	39.0	40.2	2,409(0.5%)	4,644(0.8%)	2.1e-04	1.0e-04
Coverage	$GCov(c, z)$	90.2	**95.7**	1,530(0.2%)	1,564(0.3%)	1.0e-04	5.1e-05
Long-term Relation	$G\eta(u, v, a, d)$	**9.7**	5.0	5,459(0.9%)	45,487(7.8%)	2.5e-05	1.9e-06
No memorization	$G\mu(q, e)$	98.1	**99.6**	104,467(18%)	8,577(1.5%)	1.8e-07	4.2e-07